DC Equipment Identification using K-means Clustering and kNN Classification Techniques

Abstract—**Detection of steady states and identification of small electrical loads in a household or office grid are important in efficient smart energy management. This paper proposes a method that combines two machine learning techniques, unsupervised K-means clustering, and supervised k-Nearest Neighbours classification techniques, to train a system that can effectively identify the low voltage DC electrical load, and at the same time detect whether it is in its steady state. This is done by comparing the features extracted from signatures of the electric current waveforms of equipment. The combination of K-means and kNN in the initialisation stage removes the need to know all the training elements beforehand, and thus, considerably simplifies the process. In the normal operation stage, kNN was used to identify the new unknown test element to the cluster that has the majority votes from its nearest neighbours. The centroids obtained from the K-means clustering aided in the determination of whether the system is in steady state. The method has been successfully implemented on a low voltage DC office grid, with commonly used office equipment.**

Keywords—load monitoring; DC grid; equipment identifcation; K-means clustering; kNN classification; steady state detection

I. INTRODUCTION

One of the common research interests in household and office energy management systems is load monitoring. Load monitoring involves knowing which equipment is turned on or off and whether the equipment has reached its steady state in normal operation[1]. Currently most of the research in this area revolves around AC power loads, this is due to the fact that most of the major electrical load are AC powered[2]. However, it is noticeable that there is an increasing number of DC power equipment in the office. Equipment that are traditionally powered by AC are now converting to DC; for example, the influx of LED lights replacing the traditional fluorescent lights and the DC powered fan replacing the conventional AC fan. Many modern office equipment are also coming in DC powered and require an AC-DC rectification adaptor, for example, laptops, mobile phones and LED TV.

Given that DC power supplies are getting more popular in the form of photovoltaic panels and batteries, there is a possibility of having DC office grid in the near future. Energy management systems in the future might have to deal with DC-powered loads without the intervention of AC-DC rectification process. One way to manage loads is to connect the loads with the main computer via communication means such as Ethernet lines, or wirelessly, through Wi-Fi. However, these will signify extra cost and additional modification. These add-on components can be installed in bigger equipment, such as a printer or television, but will not be cost-effective and feasible for smaller equipment such as lights and mobile phone chargers. Another alternative, as proposed in this paper, is to use a current transducer installed along the DC grid to sense the current consumption.

These electric current waveforms will form the signature of the equipment where machine learning techniques can be used to decipher, analyse, learn, and identify as required[3, 4]. Machine Learning has been used a lot in the smart energy management system, and k-Nearest Neighbours (kNN) is one of the more commonly used classification machine learning techniques in identifying and classifying the test element[5]. However, kNN requires the elements in the test space to be known and labelled before the identification process. This will be troublesome and tedious if the number of elements are large.

The proposed method in this paper uses unsupervised k-Means clustering and supervised kNN in the initialisation stage, to remove the tedious process of knowing all elements beforehand. Both methods require features to be extracted from the signature obtained from the electric current waveforms of the equipment. These features become the attributes of the training and testing elements. The unsupervised K-means clustering, groups the collection of unknown elements into their respective clusters, based on similarity in their features[6]. The clusters are then labelled using the kNN technique, by injecting a known test element into the training space. In the normal operation stage, new unknown test elements will be created from the acquired signal of the electric current waveform. These new unknown test elements will be identified to a labelled cluster using kNN algorithm by the majority votes from its nearest neighbours. The steady state of the equipment can be determined when the test element is within a distance from the centroids obtained in the K-means clustering during initialisation stage.

In this paper, a 12V DC office grid with common office equipment is set up. The proposed method was applied to the DC office grid to initialized the training set which will be used to identify the equipment and to determine when they are in steady state. The office grid comprises three 12V equipment, namely LED desk light, table top fan, and a 22” LED TV. The experiment was made more challenging by including a 12V to 19V step up DC-DC converter for laptop and a 12V to 5V step down DC-DC converter for mobile phone charging. The following section will discuss the experimental set up, the initialisation and training stage. This will be followed by the identification process in the normal operation stage and the conclusion.

II. INITIALISATION STAGE

The objective of the initialisation process is to establish a database for the identification process that will be performed during the normal operation. Fig. 1 shows the steps involved.

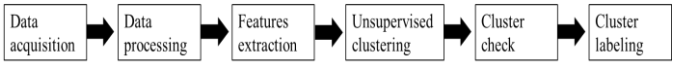


Fig. 1. Initialization stage

In this initialisation stage, both unsupervised and supervised machine learning techniques will be applied. However, the machine will still require the inputs from the users to input the number of clusters and to indicate the appliance names. This will be done via a simple text-based human machine interface.

A. *Acquisition of steady state electric current waveforms from equipment*

A 12V DC office grid is set up for this experiment. It includes five commonly-used office equipment. In addition to the 12V powered equipment of LED desk light, table top fan and 22" LED TV, the office grid also includes a 12V to 5V step down DC-DC converter for mobile phone charging and a 12V to 19V step up DC-DC converter for laptop. A current transducer was installed along the live line of the DC circuit to acquire the signal of the electric current waveform. This acquired electric current signature is sent to a computer for processing. Fifty readings of steady state electric current waveforms were acquired without labelling. Each waveform reading was made up of 300 data points acquired at a frequency of 500Hz. Fig. 2 shows all the 50 sample steady state waveforms.

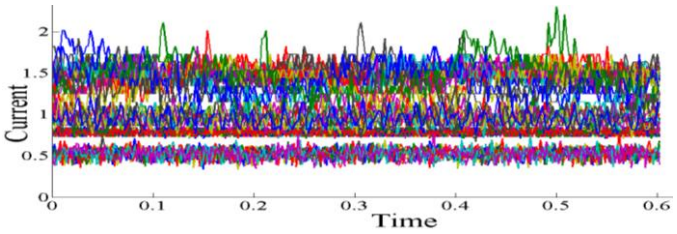


Fig. 2. Wavesforms from 5 equipment to be used in training

This experiment is interested in clustering and identifying individual office equipment in the office grid, thus only one load is active at any point of time without any overlapping signal from other equipment.

B. *Features extraction*

Ideally, the DC current consumption of electrical load should be a constant value; therefore, the simplest way to assume the operation of any equipment is by comparing the current consumption from the power source to the mean operating current of the office equipment. However, it is observed from the experiment that the DC current consumption of equipment is not always a nice straight line. On top of the Gaussian white noise in the DC line, there are other noises caused by external disturbance or due to internal components.

In detection of states and identification of loads, it is much faster and resource-efficient to use extracted features instead of using every data point from the data set. Extracted features are derived values from a data set that are informative and non-redundant. Good features should be descriptive of the data set and should help to reduce the dimensions of the data set significantly. Mean operating current value is obviously one of the most distinctive features. The mean or average value, i_μ, is the expected value of the data set which can be calculated by summing up the values of all data points divided by the size of the data set (1).

$$i_\mu = \frac{\sum_{n=1}^{m} i_n}{m} \quad \textit{where m is the data set size} \tag{1}$$

The next useful feature will be the variance of the electric current during steady state. The variance value, i_{var}, is the squared of the standard deviation value, i_σ, of the data set. The standard deviation is a measure of how spread out the data points' values are from its mean (2); and this is used to quantify the amount of variation or dispersion of a data set.

$$i_\sigma = \sqrt{i_{var}} \text{ where } i_{var} = \frac{\sum_{n=1}^{m}(i_n - i\mu)^2}{m} \tag{2}$$

Although the two extracted features, i_μ and i_{var}, in general, are able to describe most of the electric current waveforms of the office equipment, they fall short in the unsupervised clustering process. Therefore, on top of the above two familiar features, the paper uses an alternative variance, i_{avar}, as the third feature. Instead of the above mentioned variance which is the mean-square deviation from average, i_{avar} is defined as the mean of squared difference of successive data (3).

$$i_{avar} = \frac{\sum_{n=1}^{m}(i_{n+1} - i_n)^2}{m-1} \tag{3}$$

This alternative variance provides information on how fast the changing 'zigzag' pattern of the electric current waveform is during the steady state. The next section will elaborate on the use of elements, ξ, formed by the extracted features for unsupervised clustering process.

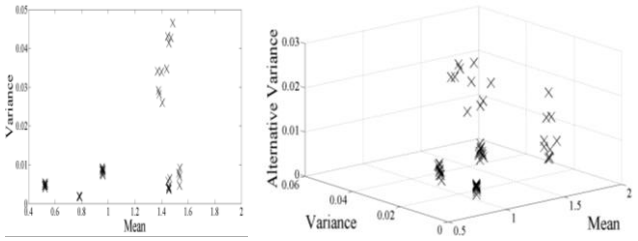
C. *Unsupervised clustering of extracted features*

K-means clustering method is an unsupervised machine learning algorithm that searches through the features of the elements in the training set and attempts to group elements with similar features together. It requires a pre-specified number of clusters, K; K = 5 in this paper, as there are five office equipment in the office grid. The elements, ξ, in the training space, $\boldsymbol{S}$, are described by the extracted features. ($\boldsymbol{S} \subseteq \mathbb{R}$; $\xi_1, \xi_2, \xi_3, \ldots, \xi_{50} \in \boldsymbol{S}$). The number of features determines the number of dimensions of the space. Fig. 3 shows the elements in the 2D and 3D spaces.

In order to group elements with similar features together, the K-means method goes through an iterative refinement process that seeks to minimise the distance between the members of each cluster and its centroid, $\dot{C}$, which is the mean of the cluster. The method starts by randomly assigning all elements in the space to one of the K groups and then calculating the error sum of squares, which is Euclidean distance, d, in 3 dimensional space (4).

$$d(x,y,z) = \sqrt{(x_\xi - x_{\hat{C}})^2 + (y_\xi - y_{\hat{C}})^2 + (z_\xi - z_{\hat{C}})^2} \quad (4)$$

A new clustering is then processed by removing each element one by one and placing them in clusters where the centroid is closest to them. This will form clusters with reduced error sum of squares. This process is repeated until convergence is reached.



(a) Elements described by 2 features (b) Elements described by 3 features

Fig. 3. Elements in 2D and 3D spaces

As the extracted features are of different scales, normalisation is required to ensure that the mean feature are not disproportionally more important than both variances during the calculation of distance. Normalised values of features are obtained by dividing the original values of an individual element's features by the features' maximum value.

One issue with the k-means clustering method is that convergence can occur at local optimum, thus global optimum is not guaranteed. If that occurs, the process will have to restart. The other issue is that if the elements in the space are not well separated, there will be incidents when the elements are wrongly assigned to another cluster. Fig. 4 shows an example of the incorrect clustering of elements. In order to check the reliability of the clusters, it is necessary to perform silhouette analysis which will be discussed in the next section.

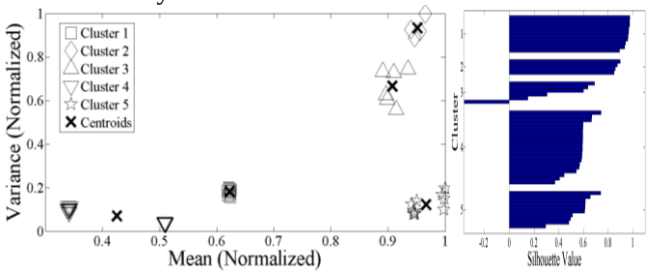


Fig. 4. Incorrect unsupervised clustering with unacceptable silhouette analysis plot

D. *Cluster reliability check using Silhouette analysis*

Silhouette analysis is a method to validate the consistency and reliability of clusters after the unsupervised K-means clustering process. It studies the separation distance between the resulting clusters and plots out a measure that shows how close an element of a cluster is to other elements in neighbouring clusters. The range of this measure is between -1 and +1. Elements with silhouette value near to +1 denote that they are far away from other clusters. Elements with silhouette value near to 0 indicate that they are very close to the decision boundary between 2 neighbouring clusters. Elements that might have been assigned incorrectly will have negative values. The K-means clustering process is considered successful when all the elements have high value. If there are elements with low or negative values, then it is necessary to restart the clustering process. If the occurrence of negative values persists after many iterations, then it implies that there are not enough features or there is an incorrect number of clusters. The silhouette plot in Fig. 4 implies that not all elements are well-grouped in the clusters, which implies that 2 features are not enough to clearly distinguish the clusters. An additional feature, i_{avar}, was added into the clustering process to form 3-dimensional clusters. Fig. 5 shows the result of a successful K-means clustering process after using 3 features.

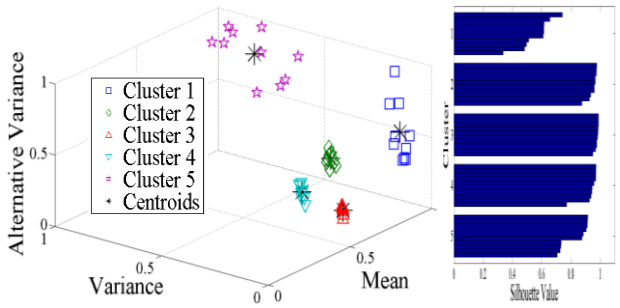


Fig. 5. Successful unsupervised clustering

This section shows that the unsupervised K-means clustering method was able to group the 50 raw steady state electric current waveforms into 5 distinct clusters by using the extracted features of i_{μ}, $i_{var,}$ and i_{avar}. However, the clusters are only labelled with cluster 1, 2, 3, 4, and 5, instead of the actual equipment names and, therefore, will not be useful in the identification process. In the next section, the supervised machine learning method, k-Nearest Neighbours, will be used to label the clusters.

E. *Supervised labelling of clusters*

KNN is a supervised machine learning algorithm that is based on the computation of the *k* nearest training elements in the overall training set and on the election of the class through majority voting on the labels of the nearest elements. It is a distance-based algorithm; this paper uses Euclidean distance to determine the closeness of the training elements to the test element. It is usually used to classify a new test element; however, in this case, it is used to label the clusters instead.

The extracted features, i_{μ}, $i_{var,}$ and i_{avar}, contribute to the attribute values of the training elements. In order for the user to label a cluster, the user will need to start the equipment in the DC office grid. After the equipment goes into steady state, the signal of the electric current waveform can be acquired by the current transducer and sent to the computer. The features of the waveform will be extracted in the same way as the training elements, in order to create the new test element. This new test element will then be thrown into the training set, and the similarities between all training elements and it will be calculated, using Euclidean distance. It will be classified to a cluster, based on majority voting of its nearest neighbours using the kNN algorithm. The user will then be prompted to

enter a label name for the test element, which is also the label name for that particular cluster. All the training elements of that cluster will be classified and labelled with the same label name. This process is repeated for all the 5 clusters in the training set. Fig. 6 shows the labelled clusters. Using this method, the user will only need to keep track of 5 different test element labels instead of all the 50 training elements. This will be significant when the training set is very large.

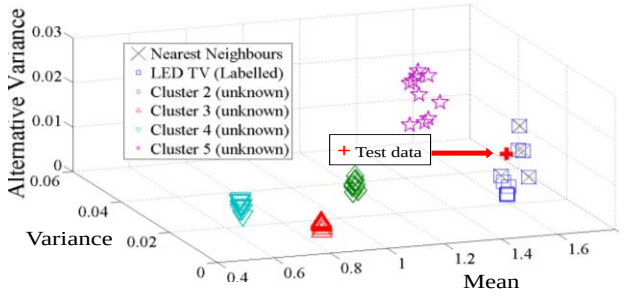


Fig. 6. Example of supervised labelling of cluster

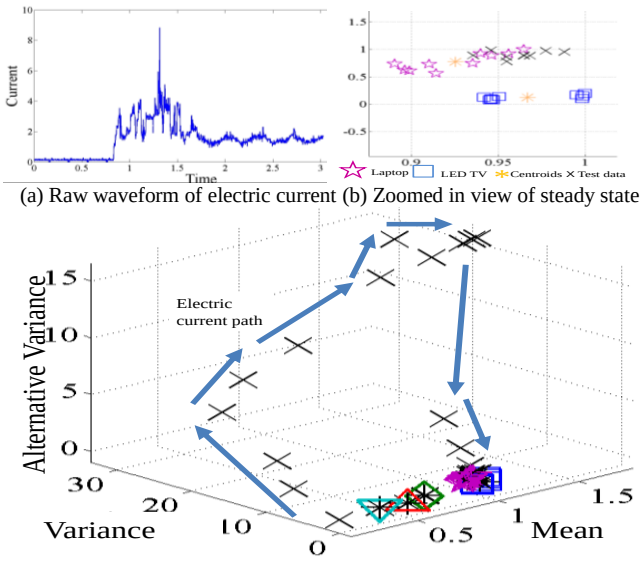
The training space will now become the testing space for the identification of equipment while detecting the state of the DC office grid during normal operation of the office grid. This will be discussed in the next section.

III. Identification of Equipment with Detection of Steady State

After the initialisation of the clusters where all clusters are well-defined and labelled, the system is ready to run in normal operation with kNN as the equipment identification algorithm. During normal operation, the data acquisition will be running constantly. The computer will process 300 data points, similar to the number of data points used in the initialisation stage, to extract the 3 selected features, namely i_{μ}, $i_{var,}$ and i_{avar}. These features will form the attributes of the test element. The test element will be placed among the trained elements in the testing space. Once again, using kNN algorithm, the test element is classified according to the 5 nearest neighbours by Euclidean distance. It is very resource-intensive to do the data processing, feature extraction and classification for every subsequent data step; therefore, the experiment did the classification after every 50 steps to reduce the workload of computer. This step size is adjustable, depending on the needs.

As the trained elements in the testing space are formed by the features that describe the steady state of individual office equipment, this method can also be used to detect steady state and transient state of the equipment. The plotting of the test element in the testing space showed the state change processes from off state to transient state to steady state.

Although a steady state of the equipment can be assumed to be established when there are more than a certain number of subsequent test elements falling into the same cluster, the detection can be further enhanced by setting a boundary around the centroid of the cluster which was determined earlier by K-means clustering. The steady state of the equipment can thus be defined, by having a certain number of subsequent test elements being identified by the same cluster, and within the boundary set around the centroid. This experiment assumed that a steady state has been reached when there are at least 5 subsequent test elements fulfilling the 2 criteria. This number can be adjusted according to needs. Fig. 7 shows the electric current path of the test element from the start of a laptop to its steady state.



(a) Raw waveform of electric current (b) Zoomed in view of steady state

(c) Electric current path of waveform

Fig. 7. Stages of an electric current upon starting up

IV. Conclusion

This paper presented an innovative way to combine unsupervised K-means clustering and supervised kNN techniques for equipment identification and steady state detection for low voltage DC equipment. The tedious process of labelling all training data for kNN classification can be removed with the intervention of K-means clustering and the boundary around K centroids enhances the steady state detection. The proposed methodology was successfully experimented in an office grid with 5 different equipment and 3 different voltages.

References

[1] S. Cao and R. R. Rhinehart, "An efficient method for on-line identification of steady state," *Journal of Process Control*, vol. 5, pp. 363-374, 1995.

[2] A. Reinhardt, D. Burkhardt, M. Zaheer, and R. Steinmetz, "Electric appliance classification based on distributed high resolution current sensing," in *Local Computer Networks Workshops (LCN Workshops), 2012 IEEE 37th Conference on*, 2012, pp. 999-1005.

[3] F. Sultanem, "Using appliance signatures for monitoring residential loads at meter panel level," *Power Delivery, IEEE Transactions on*, vol. 6, pp. 1380-1385, 1991.

[4] L. Wang, X. Luo, and W. Zhang, "Unsupervised energy disaggregation with factorial hidden Markov models based on generalized backfitting algorithm," in *TENCON 2013 - 2013 IEEE Region 10 Conference (31194)*, 2013, pp. 1-4.

[5] W. Hui, "Nearest neighbors by neighborhood counting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 942-953, 2006.

[6] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Surveys*, vol. 31, pp. 264-323, 1999.